Week 11b: Activity Key

Last Time

- Spatial EDA
- GP models to spatial data
- Spatial Prediction / Model Choice

This week

- More spatial prediction
- Anisotropic Spatial Models

Conditional Multivariate Normal Theory: Kriging

Recall, the conditional distribution, $p(Y_1|Y_2, \beta, \sigma^2, \phi, \tau^2)$ is normal with:

- $E[Y_1|Y_2] = \mu_1 + \Omega_{12}\Omega_{22}^{-1}(Y_2 - \mu_2)$

- $Var[Y_1|Y_2] = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$

Posterior Predictive Distribution

The (posterior) predictive distribution $p(Y(s_0)|y)$ can be written as

$$p(Y(s_0)|y) = \int p(Y(s_0)|y,\theta)p(\theta|y)d\theta$$

where $\theta = \{\beta, \sigma^2, \phi, \tau^2\}$.

The posterior predictive distribution gives a probabilistic forecast for the outcome of interest that does not depend on any unknown parameters.

These previous STAN code uses the Kriging idea to create a posterior predictive distribution at other locations, which results in a probabilitic prediction.

Model Evaluation for Prediction

We will use cross-validation or a test/training approach to compare predictive models.

Consider three data structures: continuous, count, and binary; how should we evaluate predictions in these situations?

Loss Functions

- Loss functions penalize predictions that deviate from the true values.

- For continuous or count data, squared error loss and absolute error loss are common.

- With binary data, a zero-one loss is frequently used.

However, these metrics are all focused on point estimates.

If we think about outcomes distributionally, empirical coverage probability can be considered. For instance, our 95 % prediction intervals should, on average, have roughly 95 % coverage.

With interval predictions, the goal is to have a concentrated predictive distribution around the outcome.

CRPS

The Continuous Rank Probability Score (CRPS) defined as

$$CRPS(F, y) = \int_{-\infty}^{\infty} \left(F(u) - 1(u \geq y)\right)^2 du,$$

where F is the CDF of the predictive distribution, is a metric that measures distance between an observed value and a distribution.

```r
crps_sample(y = 0, dat = 0)
```

```
[1] 0
```

```r
crps_sample(y = 0, dat = 2)
```

```
[1] 2
```

```r
crps_sample(y = 0, dat = c(2,-2))
```

```
[1] 1
```

```r
crps_sample(y = 0, dat = c(2,0,-2))
```

```
[1] 0.4444444
```

```r
crps_sample(y = 0, dat = c(2,0,0,0,0,-2))
```

```
[1] 0.1111111
```

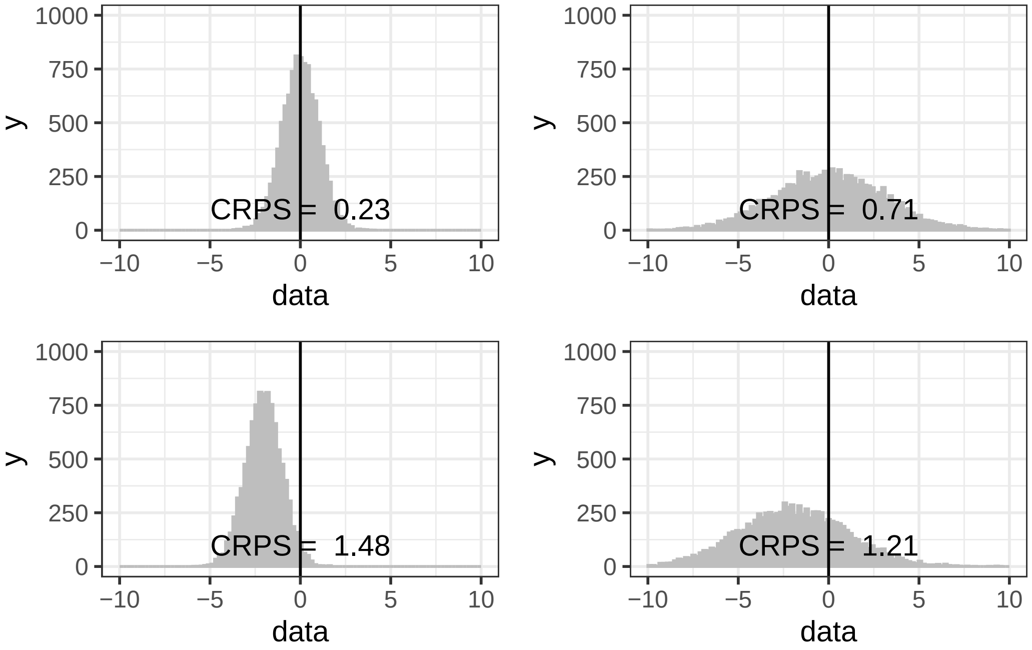
Consider four situations and sketch the predictive distribution and the resultant CRPS for each scenario. How does the MSE function in each setting?

1. Narrow predictive interval centered around outcome.

2. Wide predictive interval centered around outcome.

3. Narrow predictive interval with outcome in tail.

4. Wide predictive interval with outcome in tail.



The MSE will be quite similar for the two figures in each row, but CRPS gives us different results.

Covariance Functions

Given the assumption that a Gaussian process is reasonable for the spatial process, a valid covariance function needs to be specified.

Up to this point, we have largely worked with isotropic covariance functions. In particular, the exponential covariance functions has primarily been used. However, a Gaussian process is flexible and can use any valid covariance function.

A valid covariance function $C(h)$, needs to be a positive definite function, which includes the following properties

1. $C(0) \geq 0$
2. $|C(h)| \leq C(0)$

There are three approaches for building correlation functions. For all cases let $C_1, \ldots, C_m$ be valid correlation functions:

1. *Mixing:* $C(h) = \sum_i p_i C_i$ is also valid if $\sum_i p_i = 1$.

2. *Products:* $C(h) = \prod_i C_i$

3. *Convolution:* $C_{12}(h) = \int C_1(h-t) C_2(t) dt$ this is based on a Fourier transform.

Smoothness

Many one-parameter isotropic covariance functions will be quite similar. Another consideration for choosing the correlation function is the theoretical smoothness property of the correlation.

The Matern class of covariance functions contains a parameter, ν, to control smoothness. With $\nu = \infty$ this is a Gaussian correlation function and with $\nu = 1/2$ this results in an exponential correlation function.

"Expressed in a different way, use of the Matern covariance function as a model enables the data to inform about ν; we can learn about process smoothness despite observing the process at only a finite number of locations."

Anisotropy

Anisotropy means that the covariance function is not just a function of the distance $||h||$, but also the direction.

Geometric anisotropy refers to the case where the coordinate space is anisotropic, but can be transformed to an isotropic space.

If the differences in spatial structure are directly related to two coordinate sets (lat and long), we can create a stationary, anistropic covariance function

Let

$$cor(Y(s+h), Y(s)) = \rho_1(h_y)\rho_2(h_x),$$

where $\rho_1()$ and $\rho_2()$ are proper correlation functions.

In general consider the correlation function,

$$\rho(h;\phi) = \phi_0(||Lh||;\phi)$$

where L is a $d \times d$ matrix that controls the transformation.

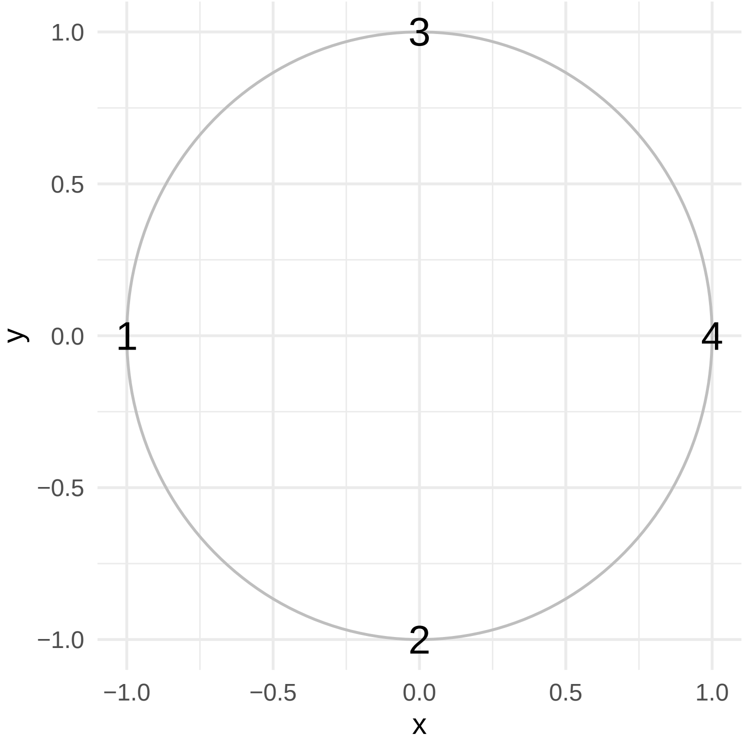
Let $Y(s) = \mu(s) + w(s) + \epsilon(s)$, and $Y(s) \sim N(\mu(s), \Sigma(\tau^2, \sigma^2, \phi, B))$, where $B = L^T L$.

The covariance matrix is defined as $\Sigma(\tau^2, \sigma^2, \phi, B)) = \tau^2 I + \sigma^2 H((h^T B h^T)^{\frac{1}{2}})$, where $H((h^T B h^T)^{\frac{1}{2}})$ has entries of $\rho((h_{ij}{}^T B h_{ij}{}^T)^{\frac{1}{2}}))$ with $\rho()$ being a valid covariance function, typically including ϕ and $h_{ij} = s_i - s_j$.

B is often referred to as a transformation matrix which rotates and scales the coordinates, such that the resulting transformation can be simplified to a distance.

Geometric Anisotropy Visual

- Consider four points positioned on a unit circle.



Now consider a set of correlation functions. For each, calculate the correlation matrix and discuss the impact of B on the correlation. Furthermore, how does B change the geometry of the correlation between points 1, 2, 3, and 4?

1. $\rho() = \exp(-{h_{ij}}^T B {h_{ij}}^T)^{\frac{1}{2}}))$, where $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

2. $\rho() = \exp(-{h_{ij}}^T B {h_{ij}}^T)^{\frac{1}{2}}))$, where $B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$

3. $\rho() = \exp(-{h_{ij}}^T B {h_{ij}}^T)^{\frac{1}{2}}))$, where $B = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}$

1. $\rho() = \exp(-{h_{ij}}^T I {h_{ij}}^T)^{\frac{1}{2}}))$

1.000	0.243	0.243	0.135
0.243	1.000	0.135	0.243
0.243	0.135	1.000	0.243
0.135	0.243	0.243	1.000

2. $\rho() = \exp(-{h_{ij}}^T B {h_{ij}}^T)^{\frac{1}{2}}))$, where $B = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$

1.000	0.177	0.177	0.059
0.177	1.000	0.135	0.177
0.177	0.135	1.000	0.177
0.059	0.177	0.177	1.000

3. $\rho() = \exp(-{h_{ij}}^T B {h_{ij}}^T)^{\frac{1}{2}}))$, where $B = \begin{pmatrix} 3 & 1 \\ 1 & 1 \end{pmatrix}$

1.000	0.243	0.086	0.031
0.243	1.000	0.135	0.086
0.086	0.135	1.000	0.243
0.031	0.086	0.243	1.000

The (effective) range for any angle η is determined by the equation

$$\rho(r_\eta({\tilde{h}_\eta}^T B {\tilde{h}_\eta}^T)^{\frac{1}{2}}) = .05,$$

where $\tilde{h}_\eta$ is a unit vector in the direction η.

Okay, so if we suspect that geometric anisotrophy is present, how do we fit the model? That is, what is necessary in estimating this model?

- In addition to σ^2 and τ^2 we need to fit B.
- While B is a matrix, it is just another unknown parameter.
- To fit a Bayesian model we need a prior distribution for B. One option for the positive definite matrix is the Wishart distribution, which is a bit like a matrix-variate gamma distribution.